DOCUMENT RESUME

ED 232 579 HE 016 503

AUTHOR Valiga, Michael J.
TITLE Assessing the Reliability of Survey Instruments. AIR 1983 Annual Forum Paper.
PUB DATE May 83
NOTE 26p.; Paper presented at the Annual Forum of the Association for
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Analysis of Variance; Higher Education; *Institutional Research; *Item Analysis; *Questionnaires; Surveys; *Test Reliability
IDENTIFIERS *AIR Forum; American College Testing Program

ABSTRACT

An analysis of variance approach to estimating reliability is examined. This approach uses an internal-consistency index for estimating the reliability of survey instruments containing ranked items, and is recommended when the relative ranking of item means is of interest to survey researchers. The computation of this index is demonstrated using data from three American College Testing Program surveys: Student Opinion Survey, Entering Student Survey, and Adult Learner Needs Assessment Survey. Eighteen of the 250 postsecondary institutions that utilized these surveys between 1980 and 1983 were studied. Survey records were analyzed using a coefficient reliability procedure available through the Statistical Package for the Social Sciences. The reliability estimate requires little computation, since the mean square values necessary to calculate the estimate may be obtained directly from analysis of variance tables. Since the approach is based on the internal-consistency of raters' responses, only one administration of the instrument is necessary. The major problem in using the proposed approach centers on the treatment of missing data. A bibliography is appended. (SW)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED232579

Assessing the Reliability of Survey Instruments

Michael J. Valiga

Program Associate

The American College Testing Program

P.O. Box 168

Iowa City, Iowa 52243

319/337-1102

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☐ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

ATP

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

IHE 016 503

Running head: Assessing the Reliability of Survey Instruments

THE ASSOCIATION FOR INSTITUTIONAL RESEARCH

This paper was presented at the Twenty-Third Annual Forum of the Association for Institutional Research held at the Sheraton Centre in Toronto, Ontario, May 23-26, 1983. This paper was reviewed by the AIR Forum Publications Committee and was judged to be of high quality and of interest to others concerned with the research of higher education. It has therefore been selected to be included in the ERIC Collection of Forum papers.

D. R. Coleman, Chairman
Forum Publication
Advisory Committee

Abstract

In recent years, college personnel have made increasing usage of surveys and questionnaires to collect a variety of data to aid in administrative decision making. Due to this increased usage, the reliability of these instruments has become an issue of interest at many postsecondary institutions. Although numerous approaches to assessing instrument reliability currently exist, commonly-used internal-consistency reliability indices (such as KR_{20}, KR_{21}, and coefficient α) may not be appropriate for certain uses made of survey instruments, particularly in the case in which the relative magnitudes of mean item ratings (across all respondents) are of interest. This paper briefly examines current approaches used for estimating survey reliability, and presents an alternative approach based on the variability of mean item ratings. An Analysis of Variance (ANOVA) approach is utilized to compute the reliability estimate. Due to the relative ease with which this approach may be used, the method is suggested for documenting survey reliability at individual institutions.

Assessing the Reliability of Survey Instruments

During recent years, an increasing variety of measurement instruments have been constructed to provide data for decisions which affect the educational, political and social affairs of nearly everyone in society. In particular, the questionnaire or survey instrument has gained widespread acceptance as a tool for collecting psychometric data in a variety of settings. Between 1970 and 1976, for example, six federal agencies alone commissioned more than 200 opinion- and attitude-related studies (Carter, 1979). A similar burgeoning of survey and polling research has occurred in the private sector, where market research has become an increasingly sophisticated and commonplace endeavor (Aaker and Day, 1980). While it is difficult to estimate the degree to which educational institutions utilize surveys and questionnaires, an examination of education-related literature indicates that this utilization is also increasing.

Due to this increasing utilization, many issues and questions have been raised concerning the quality, accuracy and utility of survey data. As is the case with all measurement instruments, it is the responsibility of the developer of each survey instrument to address these issues, and to provide adequate technical data to assist potential users in the selection and use of the instrument.

One of the primary types of technical information presented by producers of psychometric instruments deals with the concept of instrument reliability. Nunnally (1959, p. 95) defines reliability as "the precision of measurement regardless of what is being measured". A more statistically oriented definition is given by the American Psychological Association (1974, p. 48):

"Reliability is the degree to which the results of measurement are attributable to systematic sources of variance ("true" variance)".

As is the case with all psychometric instruments, the reliability of survey instruments is typically estimated using one of two general approaches. The first of these approaches, commonly referred to as test-retest reliability, repeated measures reliability, or parallel-forms reliability, involves two administrations of an instrument (or of parallel forms of an instrument) and an examination of the stability of the results obtained. Using this approach, the statistic of interest (individual total score, percent "positive", group mean, etc.) is calculated for all cases (or groups) in both of the samples tested or surveyed. The estimate of reliability is then determined by computing one of several measures of association between the two sets of statistics. Measures of association which have frequently been used to describe the reliability of survey instruments include the percent of similar responses (the American College Testing Program, 1981; Havighurst, 1970; and van Es and Wilkening, 1979), the χ^2 statistic (Pucel and Luftig, 1975), the Kappa coefficient (Crittenden and Hill, 1971; Gottfredson and Holland, 1975; and Lawlis and Lu, 1972), and the Pearson product-moment correlation coefficient (Eagle, 1977; Harris, 1979; Marsh and Overall, 1979b; and Overall and Marsh, 1980). Other measures of association occasionally used by survey researchers include Spearman's rank-order correlation coefficient (rho) and the correlational coefficient (e) proposed by Jobsen (1976).

The second general approach to estimating survey reliability involves a single administration of an instrument and the computation of some type of internal-consistency index based on the data. This approach assumes that the instrument contains a series of items related to a particular topic of interest

and that the item scores may be summed to yield a "total" scale score for each individual (or object of measurement). Reliability is subsequently defined in terms of the internal consistency of the items in the scale. Internal-consistency indices which have often been used in studying survey reliability include the "split-halves" and odd-even correlational approach (Eagle, 1977; and Karstensson, 1972), the Kuder-Richardson formulas KR_{20} and KR_{21} (Kahn and Roberts, 1971) and Cronbach's coefficient α (Marsh and Overall, 1979a; Masters, 1974; and Miskel and Heller, 1972).

Several variations of the two general approaches for determining instrument reliability have also been employed by survey researchers. These variations include several approaches involving interrater reliability (Deshpande, Webb and Marks, 1970; Feldman, 1977; Gillmore, 1973; Menne, 1968; Murray, 1975; and Remmers and Wersbrodt, 1964), techniques employing analysis of variance (ANOVA) results (Centra, 1972; Centra, 1973; Marsh, 1977; and Marsh and Overall, 1979b), and more recent approaches using generalizability theory (Carbno, 1981; Gillmore, Kane and Naccarato, 1976; Kand and Brennan, 1977; and Kane, Gillmore and Crooks, 1976).

In nearly all of the previously cited studies, the estimates of survey reliability were based on individual scores or on class means, teacher means or school means (derived from individual scores). In these cases, the variability among the individual scores or group means was a major determinant of the magnitude of the reliability index. Survey researchers, however, are often confronted with situations in which individual scores are of limited interest, while the relative magnitudes of the mean item ratings (across respondents) are of primary importance. For example, college personnel may be interested in

comparing the average satisfaction ratings assigned to a series of college services by students at the institution. As a second example, secondary school administrators may wish to prioritize their guidance objectives on the basis of average student responses to a survey containing a list of potential needs of high school students. In both examples, the primary interest is in the mean ratings of the survey items, and therefore, the index of reliability should address the stability/consistency of these mean ratings. Unfortunately, survey researchers typically present "traditional" reliability estimates based on individual scores regardless of the uses made of the survey data. Although a limited number of survey-related studies have examined the reliability of mean item ratings (Brittingham and Netusil, 1976; Eagle, 1977; Hicks, 1970; and Parry-Hill, 1975), the techniques used in these studies are not well known and are rarely applied by survey researchers, even when such applications would be appropriate.

The purpose of this paper is to present an easy-to-use index for estimating survey reliability based on the internal consistency of mean item ratings, and to demonstrate the computation of this index using survey data available through the ACT Evaluation/Survey Service.

Method

The statistical method for estimating reliability presented in this paper is based on the analysis of variance approach to interrater reliability discussed by Ebel (1951) and Winer (1962 and 1971). Indeed, this method may be thought of as a special case of interrater reliability, in which raters rate the items of the survey instrument rather than persons, classes, etc. The development presented in this section, therefore, closely parallels that presented in Winer's statistical texts.

Conceptually, this approach is similar to the "standard" internal-consistency estimates of reliability in which the variability of individual total scores (summed across items) forms the basis for the reliability coefficient. The approach proposed in this paper involves the same person x item matrix, however, the interest is in the relative magnitudes of mean item scores (summed across persons). The variability among the item means, therefore, serves as the basis for the estimate of reliability. Because the proposed coefficient is based on the variability of item means rather than on individual score variability, the proposed coefficient could be obtained by submitting an items x persons data matrix (rather than a persons x items matrix) to a standard computer program which computes Cronbach's coefficient α. It is possible, however, to obtain the reliability estimate directly from the mean square values provided in a standard ANOVA table (provided by many reliability programs which compute coefficient α) without transposing the persons x items matrix.

The Model

A given item j may be assumed to possess a "true" magnitude (τ) of some particular characteristic (level of student satisfaction, level of student need, etc.). In measuring this characteristic with some psychometric instrument administered to i individuals, the observed score may be represented as:

$$X_{ij} = \tau_j + \pi_i + e_{ij}$$

where X_{ij} = the observed measurement

τ_j = the true magnitude of the characteristic being measured for item j

π_i = the main effect (anchor point) for the ith rater

e_{ij} = error of measurement.

Upon repeated measurement with the same instrument (i.e., similar raters), τ_j is assumed to remain constant, while π_i (for raters) and e_{ij} are assumed to vary. The mean ($\bar{I}_j$) for a given item is defined as:

$$\frac{\sum_{i=1}^{n} X_{ij}}{n} = \bar{I}_j = \tau_j + \bar{\pi}. + \bar{e}._j$$

The general definition for the reliability of a particular series of measures is:

$$\rho = \frac{\text{true variance of the measures}}{\text{true variance of the measures + error variance for the model}}$$

In the case of the model listed above, if the differences due to anchor points (raters) are not considered part of the error of measurement, the reliability of the item means ($\bar{I}_j$'s) is given by:

$$\rho_{\text{Item Means}} = \frac{\sigma_\tau^2}{\sigma_\tau^2 + \sigma_{\bar{e}}^2} = \frac{\sigma_\tau^2}{\sigma_\tau^2 + \sigma_{\frac{e}{n}}^2}$$

To estimate this reliability, therefore, it is only necessary to obtain estimates of σ_τ^2 and σ_e^2. These estimates may be derived from an Analysis of Variance for the model outlined earlier (using an items x raters matrix):

Source of Variation	SS	df	MS	E(MS)
Between Items	$SS_{b.\ items}$	(k-1)	$MS_{b.\ items}$	$\sigma_e^2 + n\sigma_\tau^2$
Within Items	$SS_{w.\ items}$	k(n-1)	--	--
Between Raters	$SS_{b.\ raters}$	(n-1)	$MS_{b.\ raters}$	$\sigma_e^2 + k\sigma_\pi^2$
Residual	SS residual	(n-1)(k-1)	MS residual	σ_e^2

Note that σ_e^2 may be estimated directly using MS_{resid}, and that $MS_{b.\ Items} = \sigma_e^2 + n\sigma_\tau^2$. Therefore,

$$\hat{\sigma}_\tau^2 = \frac{MS_{b.\ Items} - MS_{resid.}}{n}$$

Using these two estimates, the estimated reliability coefficient may be determined:

$$\hat{\rho} = \frac{\frac{MS_{b.\ Items} - MS_{resid}}{n}}{\frac{MS_{b.\ Items} - MS_{resid}}{n} + \frac{MS_{resid}}{n}}$$

$$= \frac{MS_{b.\ Items} - MS_{resid}}{MS_{b.\ Items}}$$

$$= 1 - \frac{MS_{resid}}{MS_{b.\ Items}} \qquad (1)$$

Note that to compute this coefficient, it is only necessary to use two mean square values obtained from an ANOVA based on the items x persons data matrix.

The mean square values necessary to determine this reliability estimate ($MS_{b.\ items}$ and MS_{resid}) may also be obtained as by-products from an ANOVA table for the "standard" reliability situation in which the reliability of individual scores is of interest. In this type of situation, an analysis of a persons x items matrix yields the following results:

Source of Variation	SS	df	MS	E(MS)
Between People	$SS_{b.\ people}$	(n-1)	$MS_{b.\ people}$	$\sigma_e^2 + k\sigma_\pi^2$
Within People	$SS_{w.\ people}$	n(k-1)	--	--
Between Items	$SS_{b.\ items}$	(k-1)	$MS_{b.\ items}$	$\sigma_e^2 + n\sigma_\tau^2$
Residual	SS residual	(k-1)(n-1)	MS_{resid}	σ_e^2

The advantage of using this approach to obtain $MS_{b.\ items}$ and MS_{resid} is that it does not require the researcher to transpose the standard persons x items matrix prior to analyzing the data. It is interesting to note that the ANOVA formula for Cronbach's coefficient α differs from the proposed estimate only in the mean square value used in the denominator of the fraction ($MS_{b.\ people}$ as opposed to $MS_{b.\ items}$):

$$\alpha = 1 - \frac{MS_{resid}}{MS_{b.\ people}} .$$

The assumptions necessary to estimate the proposed reliability index are as follows (derived from those listed by Winer, 1962):

1. Errors of measurement are not correlated with the "true" score for each item.
2. The sample of k items on which the measurements are being made is a random sample from a population of comparable items to which inferences are to be drawn.
3. The sample of n individuals (raters) is a random sample from the population of all such individuals (raters).
4. The within-item variance may be pooled to provide an estimate of σ_e^2.

The reliability estimate outlined in this paper is based on the n individual ratings used in computing the mean square values. Reliability estimates for larger or smaller groups of raters may be obtained using the Spearman-Brown Prophecy Formula.

Instrumentation

Data for the study were obtained using the American College Testing Program (ACT) Student Opinion Survey, Entering Student Survey and Adult Learner

Needs Assessment Survey. These instruments are four-page optically scannable questionnaires containing items related to college experiences, college plans and preferences, and the needs of adult learners entering college. Each of the instruments is composed of several sections of items, including sections dealing with demographic information and a section for additional questions designed by college personnel. Without the additional questions, each instrument requires approximately 20 minutes to complete. These particular instruments were selected for the study since they are widely utilized by colleges and universities, and because each instrument contains at least one section in which the relative ratings of the items are of interest.

Section III of the Student Opinion Survey contains a list of 42 general aspects of the college environment which each student is asked to rate using a five-point Likert satisfaction scale ("Very Satisfied" through "Very Dissatisfied). Section III of the Adult Learner Needs Assessment Survey contains a list of 66 personal and educational needs. Each adult learner completing the instrument is asked to indicate the amount of assistance he or she requires in each of the need areas ("a lot of help", "a medium amount of help", "a little help", or "no further help"). Section III-A of the Entering Student Survey contains a list of 23 factors related to the college selection. Each student is asked to indicate whether each factor was "Very Important", "Moderately Important", "Slightly Important" or "Not Important" in his or her decision to attend the particular institution. In all three of these sections, the relative magnitudes of the average item ratings are typically of interest to college decision makers.

Participants

Between the fall of 1980 and the spring of 1983 approximately 250 post-secondary institutions utilized one or more of the three instruments employed in this study. These institutions were not selected in a random manner; all institutions voluntarily purchased the survey instruments and related evaluation services through the ACT Evaluation/Survey Service. The survey instruments were administered by college personnel to varying subgroups of students and potential students, using different administration modes at each institution. The Student Opinion Survey was administered primarily to enrolled undergraduate students, the Entering Student Survey to students accepted for admission and to first-year students, and the Adult Learner Needs Assessment Survey to nontraditional students planning to enter college or currently enrolled at the institution. After the instruments were completed, they were returned to ACT for processing and analysis. In this manner, data files for the three instruments containing approximately 100,000 student records were obtained.

For each of the three surveys included in the study, six institutions (three public and three private) were randomly selected from all participating institutions which administered between 250 and 700 copies of the particular instrument. The eighteen institutions identified in this manner included both large and small, and two-year and four-year institutions from across the United States.

Procedures

The mean square values necessary to compute the reliability estimate discussed in this paper were obtained using the coefficient α reliability procedure available through the Statistical Package for the Social Sciences (SPSS). The person x item matrix for each of the 18 institutions was analyzed

separately and a reliability estimate was calculated using the proposed approach. The mean square values necessary to compute the reliability estimate ($MS_{residual}$ and $MS_{b.\ items}$) were obtained directly from the analysis of variance table for the standard α reliability coefficient.

Missing data were dealt with using two different approaches. For the Student Opinion Survey, individuals that failed to answer all 42 items were deleted from the analysis. (This approach is the default option used in the SPSS reliability procedure.) For the Entering Student Survey and the Adult Learner Needs Assessment Survey two different analyses were conducted for each institution: one in which all individuals with missing data were deleted, and one in which missing values were equated to the "no need" or "not important" response. In all cases in which data were deleted, a conservative reliability estimate was obtained due to the reduced number of cases analyzed.

Results

Figure 1 contains a detailed example of the computation of the proposed "item-mean" reliability index for one of the institutions included in the study. For this institution, a total of 595 Adult Learner Needs Assessment Survey records were available for analysis, however, 81 records were eliminated because they contained missing data for one or more of the 66 need items. The remaining 514 records were analyzed using the coefficient α reliability procedure available through the SPSS statistical package. The mean square values required to compute the reliability index were obtained from the third and fourth rows of the analysis of variance table generated using this procedure. The reliability estimate obtained in this manner (based on 514 cases) was found to be .9846. Figure 1 also contains an example of the use of the

Figure 1

Computations of the item-mean reliability index for a sample institution that administered the Adult Learner Needs Assessment Survey:

Number of survey responderts = 595

Number of respondents with complete data = 514

Number of survey items = 66

Analysis of Variance table generated by the SPSS coefficient α reliability procedure:

Source of Variance	SS	df	MS
Between People	9024.29713	513	17.59122
Within People	23015.30303	33410	0.68887
Between Measures	2589.57682	65	39.83964
Residual	20425.72621	33345	0.61256
Total	32039.60017	33923	0.94448

Item-mean reliability estimate (based on 514 cases):

$$\hat{\rho}_{514} = 1 - \frac{MS_{residual}}{MS_{b.\ measures\ (items)}}$$

$$= 1 - \frac{.61256}{39.83964}$$

$$= .9846$$

Spearman-Brown estimate of the reliability index for 250 raters:

$$\hat{\rho}_{250} = \frac{\left(\frac{250}{514}\right)\hat{\rho}_{514}}{1 + \left(\frac{250}{514} - 1\right)\hat{\rho}_{514}}$$

$$= .9688$$

Spearman-Brown formula to estimate the reliability index for a larger or smaller number of raters (student respondents). Using this formula, the estimated reliability index for a sample of 250 raters was found to be .9633.

Table 1 presents the item-mean reliability estimates obtained for the 18 colleges and universities included in the study. Separate sections of the table summarize the results for each of three instruments examined.

Reliability estimates for the six schools that used the Student Opinion Survey ranged from .9010 to .9773. Since only cases with valid responses to all 42 items were included in the analysis, the number of records analyzed for several of the colleges was very limited (e.g., 58 and 92 cases). The percentage of cases eliminated during the analysis ranged from just under 50% to nearly 80%. Due to this loss of data, conservative estimates of the item-mean reliability were obtained.

Two separate analyses were conducted for each of the colleges that used the Entering Student Survey. In one of these analyses, cases with missing data were eliminated, and in the other, all missing information was coded to the "not important" response. The results of these two analyses for each college (presented in Table 1) were nearly identical, since few cases with missing data were encountered. The largest difference observed in the reliability estimates obtained using the two approaches to handling missing data was .0002. The item-rank reliability estimates obtained in these analyses ranged from .9909 to .9974.

As was the case for the Entering Student Survey, two analyses were conducted for each institution that used the Adult Learner Needs Assessment Survey--one in which missing data were eliminated and one in which missing

Table 1

Reliability Estimates for Eighteen Postsecondary Institutions Based on Mean Item Ratings

Student Opinion Survey--Section III
(Cases with missing data eliminated)

		Total Number of Cases	Number Included in Analysis	Estimate of Reliability
Public Colleges:	1.	535	270	.9629
	2.	384	124	.9423
	3.	250	58	.9010
Private Colleges:	1.	610	269	.9773
	2.	440	225	.9754
	3.	398	92	.9569

Entering Student Survey--Section III-A
(Cases with missing data eliminated)

		Total Number of Cases	Number Included in Analysis	Estimate of Reliability
Public Colleges:	1.	643	613	.9974
	2.	529	488	.9944
	3.	442	416	.9929
Private Colleges:	1.	464	434	.9945
	2.	451	416	.9960
	3.	389	374	.9909

Entering Student Survey--Section III-A
(Missing cases coded to "not important")

		Total Number of Cases	Estimate of Reliability
Public Colleges:	1.	643	.9974
	2.	529	.9945
	3.	442	.9930
Private Colleges:	1.	464	.9946
	2.	451	.9961
	3.	389	.9911

Table 1 (continued)

Adult Learner Needs Assessment Survey--Section III
(Cases with missing data eliminated)

		Total Number of Cases	Number Included in Analysis	Estimate of Reliability
Public Colleges:	1.	595	514	.9846
	2.	449	391	.9891
	3.	369	307	.9813
Private Colleges:	1.	560	489	.9906
	2.	364	303	.9878
	3.	336	196	.9806

Adult Learner Needs Assessment Survey--Section III
(Missing cases coded to "no further help needed")

		Total Number of Cases	Estimate of Reliability
Public Colleges:	1.	595	.9860
	2.	449	.9905
	3.	369	.9835
Private Colleges:	1.	560	.9913
	2.	364	.9895
	3.	336	.9870

information was coded to the "no further help needed" response. In these analyses, only one of the six institutions was found to have missing data in over 20% of its cases, and for this institution the loss was under 50%. For this reason, the results of the two analyses yielded similar results for all institutions examined. Observed reliability estimates ranged from .9806 through .9913.

Discussion and Conclusions

In situations in which the relative magnitudes of item-means are of interest to survey researchers, the approach to estimating reliability presented in this paper is both appropriate and easy to use. The reliability estimate requires little computation on the part of the researcher, since the mean square values necessary to calculate the estimate may be obtained directly from analysis of variance tables produced by reliability routines available through common computerized statistical packages (e.g., SPSS and SAS). Since the approach is based on the internal consistency of raters' responses, only one administration of the instrument is required. In addition, the reliability estimate is not difficult to interpret due to its similarity to Cronbach's coefficient α.

The major problem with using the approach outlined in this paper centers on the treatment of missing data. If all student (respondent) records with missing data for one or more of the items are eliminated, a large portion of the data may be lost, and a relatively conservative reliability estimate will be obtained. This was the case in the current study, for the six colleges which used the Student Opinion Survey. The alternative approach used for the Entering Student Survey and the Adult Learner Needs Assessment Survey involved recoding all missing data to the "not important" or "no need" response. This

approach is only appropriate in situations in which blanks may be considered as "0" responses. Other techniques which might be used to address the problem of missing data include the following:

- Any item which is left blank by a "large" portion of the survey respondents could be eliminated from the analysis.
- The item mean (based on all individuals who did respond to the item) could be substituted for all missing ratings for the particular item.
- A simple one-way analysis of variance model could be used (as suggested by Ebel, 1951), rather than the individuals x items design employed in this study.

Further research is needed into the relative merits of each of these approaches to addressing the problem of missing data.

Regardless of the manner in which missing data was handled, very high reliability estimates were obtained for the three survey instruments examined. For all 18 colleges included in the study, these estimates were above .90 and all but two of the estimates were above .95. Based on this information, it would appear that relative ratings of item means obtained using these instruments are highly reliable (given the sample sizes utilized).

Because of the ease with which the reliability index discussed in this paper may be computed and interpreted, it is strongly recommended for use (in appropriate situations) at the individual college campus.

References

Aaker, D. A., & Day, G. S. Marketing research: private and public sector decisions. New York: John Wiley, 1980.

American College Testing Program. The ACT Evaluation/Survey Service: User's Guide (2nd ed.). Iowa City, IA: ACT Publications, 1981.

American Psychological Association. Standards for Educational and Psychological Tests and Manuals. Washington, DC: American Psychological Association, 1974.

Brittingham, B. E., & Necusil, A. J. The reliability of goal ratings in a needs assessment procedure. Journal of Educational Psychology, 1976, 69.

Carbno, W. C. Student evaluation of teacher effectiveness: a case study. Educational and Psychological Measurement, 1981, 41.

Carter, M. N. Case study of an opinion poll: skill, luck, and insight rate high. Minneapolis Star, October 24, 1979.

Centra, J. A. The utility of student ratings for instructional improvement: I. the effectiveness of student feedback in modifying college instruction; II. self-ratings of college teachers: a comparison with student ratings. Princeton, NJ: Educational Testing Service, 1972.

Centra, J. A. The relationship between student and alumni ratings of teachers. Princeton, NJ: Educational Testing Service, 1973.

Crittenden, K. S., & Hill, R. J. Coding reliability and validity of interview data. American Sociological Review, 1971, 36, 1073-1080.

Cronbach, L. J. Coefficient alpha and the internal structure of tests. Psychometrika, 1951, 16.

Deshpande, A. S., Webb, S. C., & Marks, E. Student perceptions of engineering instructor behaviors and their relationships to the evaluation of instructors and courses. American Educational Research Journal, 1970, 7.

Eagle, N. Studies in the reliability and validity of a community college student rating of instruction questionnaire. Community/Junior College Research Quarterly, 197[illegible].

Ebel, R. L. Estimation of the reliability of ratings. Psychometrika, 1951, 16.

Feldman, K. A. Consistency and variability among college students in rating their teachers and courses: a review and analysis. Research in Higher Education, 1977, 6.

Gillmore, G. M. Estimates of reliability coefficients for items and subscales of the Illinois course evaluation questionnaire. (Research Report No. 341). Urbana-Champaign, IL: University of Illinois, Measurement and Research Division, Office of Instructional Resources, 1973.

Gillmore, G. M., Kane, M. T., & Naccarato, R. W. The generalizability of student instructional ratings: general theory and application to the Washington Instructional Assessment System. Seattle, WA: University of Washington, Educational Assessment Center, 1976.

Gottfredson, G. D., & Holland, J. L. Vocational choices of men and women: a comparison of predictors from the self-directed search. Journal of Counseling Psychology, 1975, 22, 28-34.

Harris, L. J. Vocational Education R & D Products: Impact, Reliability, and Validity. Columbus, OH: Ohio State University, National Center for Research in Vocational Education, 1979.

Havighurst, R. J. The reliability of rating scales used in analyzing interviews with parents, students, teachers, and community leaders. Chicago, IL: National Study of American Indian Education, Series IV, No. 9, 1970.

Hicks, R. E. A note on the test-retest reliability of a Zambian occupational prestige questionnaire. Measurement and Evaluation in Guidance, 1970, 3.

Jobson, J. D. A coefficient of equality for questionnaire items with interval scales. Educational and Psychological Measurement, 1976, 36.

Kane, M. T., Gillmore, G. M., & Crooks, T. J. Student evaluations of teaching: the generalizability of class means. Journal of Educational Measurement, 1976, 13.

Kane, M. T., & Brennan, R. L. The generalizability of class means. Review of Educational Research, 1977, 47.

Karstensson, L. A study on the validity and reliability of a questionnaire on student attitude toward economics. Athens, OH: Ohio University, Dept. of Economic Education, 1972.

Khan, S. B., & Roberts, D. M. Factorial stability of academically relevent affective characteristics. Measurement and Evaluation in Guidance, 1971, 3.

Kuder, G. F., & Richardson, M. W. The theory of the estimation of test reliability. Psychometrika, 1937, 2.

Lawlis, G. F., & Lu, E. Judgments of counseling process: Reliability, agreement, and error. Psychological Bulletin, 1972, 78.

Marsh, H. W. The validity of students' evaluations: Classroom evaluations of instructors independently nominated as best and worst teachers by graduating seniors. American Educational Research Journal, 1977, 14.

Marsh, H. W., & Overall, J. U. Validity of students' evaluations of teaching: A comparison with instructor self evaluations by teaching assistants, undergraduate faculty and graduate faculty. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, 1979a.

Marsh, H. W., & Overall, J. U. Long-term stability of students' evaluations: A note on Feldman's "Consistency and variability among college students in rating their teachers and courses". Research in Higher Education, 1979b, 10.

Masters, J. R. The relationship between number of response categories and reliability of Likert-type questionnaires. Journal of Educational Measurement, 1974, 11.

Menne, J. Students' evaluation of instructors. Presented at the Annual Meeting of the National Council on Measurement in Education, 1968.

Miskel, C., & Heller, L. E. The stability and reliability of a modified work components study questionnaire in the educational organization. Paper presented at American Educational Research Association Annual Meeting, Chicago, 1972.

Murray, H. G. Predicting student ratings of college teaching from peer ratings of personality traits. Teaching of Psychology, 1975, 2.

Nunnally, J. C. Tests and Measurements. New York: McGraw-Hill, 1959.

Overall, J. U., & Marsh, H. W. Students' evaluations of instruction: A longitudinal study of their stability. Journal of Educational Psychology, 1980, 72.

Parry-Hill, J. W., Jr. The initial development and validation of a short form occupational analysis inventory based on systematically derived work dimensions. Toward an educational system articulated with adult society. Raleigh, NC: North Carolina State University, Center for Occupational Education, 1975.

Pucel, D. J., & Luftig, J. T. The reliability of the Minnesota vocational follow-up student questionnaire. Minneapolis, MN: Dept. of Vocational and Technical Education, 1975.

Remmers, H. H., & Weisbrodt, J. A. Manual of instructions for Purdue rating scale of instruction. Purdue, IN: Purdue Research Foundation, 1964.

van Es, J. C., & Wilkening, E. A. Response stability in survey research: A cross-cultural comparison. Rural Sociology, 1970, 35.

Winer, B. J. Statistical principles in experimental design (1st ed.). New York: McGraw-Hill, 1962.

Winer, B. J. Statistical principles in experimental design (2nd ed.). New York: McGraw-Hill, 1971.